FACULTADE DE MATEMÁTICAS

Trabajo Fin de Grado

Estabilización de la varianza

Alma García Pérez

2018/2019

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Trabajo Fin de Grado

Estabilización de la varianza

Alma García Pérez

09/2019

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación operativa.

Título: Estabilización de la varianza.

Breve descripción del contenido:

Con objeto de construir intervalos de confianza para un parámetro, es frecuente emplear un método pivotal aproximado, que consiste en tomar un estimador del parámetro que, una vez estandarizado dividiendo por su error típico, converge en distribución a una normal estándar. El intervalo resultante está centrado en el estimador y su radio es el cuantil correspondiente de la normal, multiplicado por el error típico. El problema surge cuando el error típico depende del parámetro que se quiere estimar. Esto es muy frecuente. Ocurre, por ejemplo, al construir un intervalo de confianza para una proporción, o para el coeficiente de correlación, o para el parámetro de la distribución de Poisson.

Afortunadamente, en algunos casos es posible encontrar una transformación del estimador cuya desviación típica ya no dependa del parámetro en cuestión. Así se evita tener que estimar el error típico, e incluso suele obtenerse una mejor aproximación por la distribución normal. Al deshacerse la transformación, se consigue el intervalo de confianza para el parámetro original.

En este trabajo revisaremos estos métodos, que se conocen como transformaciones que estabilizan la varianza, aplicados a los ejemplos ya mencionados y se incluirán los argumentos de probabilidad que permitan obtener las propiedades de cada método estadístico, así como experimentos con datos simulados.

Índice general

Resumen viii

Introducción xi

1. Aproximación en distribución 1

1.1. Conceptos básicos . . . 1

1.1.1. Convergencia de sucesiones de variables aleatorias . . . 2

1.2. Estimación . . . 3

1.2.1. Estimación puntual . . . 3

1.2.2. Intervalos de confianza . . . 5

1.2.3. Método de la cantidad pivotal . . . 6

1.3. Teorema central del límite . . . 8

1.3.1. Aplicación a la proporción poblacional . . . 10

2. Método de estabilización de varianza 13

2.1. Método delta . . . 13

2.2. Transformación estabilizadora de la varianza . . . 14

2.2.1. Aplicación a la proporción poblacional . . . 15

2.3. Ejemplos notables . . . 17

2.3.1. Distribución de Bernoulli . . . 17

2.3.2. Distribución de Poisson . . . 18

2.3.3. Distribución exponencial . . . 19

2.3.4. Coeficiente de correlación . . . 21

3. Estudios de simulación 23

3.1. Distribución de Bernoulli . . . 23

3.2. Distribución de Poisson . . . 27

3.3. Distribución Exponencial . . . 31

3.4. Coeficiente de correlación 32

Apéndice A : Código de R **37**

Bibliografía **43**

Resumen

En el contexto de la construcción de intervalos de confianza para un parámetro desconocido, es frecuente enfrentarse al problema de no poder construirlo debido a que el error típico del estimador depende de dicho parámetro. Para solventar esta dificultad se estudiarán dos métodos: el que denominaremos usual, que se basa en la estimación del error típico y el de la transformación estabilizadora de la varianza, que es una aplicación del método delta.

Se hará un estudio de simulación para la comparación de los intervalos de confianza obtenidos por ambos métodos, en los que analizaremos las coberturas obtenidas para tres tamaños muestrales, los tres niveles de confianza usuales y distintos valores de los verdaderos parámetros. Además se añadirán distintas representaciones gráficas que muestran que con el método de la transformación estabilizadora de la varianza se suele obtener una mejor aproximación por la distribución normal y se suele mejorar la simetría de la distribución.

Abstract

In the context of the construction of confidence intervals for an unknown parameter, it is frequent to face the problem of not being able to constuct said confidence interval due to the fact that standard error of the estimator depends on that parameter. In order to solve such a difficulty, two methods will be studied: the one we will refer to as usual, which is based on the estimations of the standard error, and that of the variance stabilizing transformation, which is an application of the delta method.

A simulation study will be conducted for the comparison of the confidence intervals obtained with both methods, in which we will analize the coverages obtained for three sample sizes, the three usual levels of confidence and different values of the true parameters. Moreover, different graphic representations will be added that show that by using the variance-stabilizing transformation, a better approximation for the normal distribution is obtained. Additionally, the symetry of the distribution is often improved.

Introducción

Este trabajo tiene por objetivo estudiar las transformaciones que estabilizan varianza para la construcción de intervalos de confianza para un parámetro θ, ya que en Estadística la distribución de los estimadores es un problema importante y dicha transformación solventa el problema de que la distribución del estimador $\hat{\theta}$ dependa del parámetro a estimar. Supongamos por ejemplo, que queremos construir un intervalo de confianza para para la proporción de elementos p de una población que posee una determinada característica de interés, a partir de una muestra aleatoria simple de elementos de la población. Sabemos, por tanto que la muestra $X_1, \ldots, X_n$ es de variables independientes e idénticamente distribuidas a una variable $X \in Be(p)$ con media p y varianza $p(1-p)$.

Para la construcción de intervalos de confianza para un parámetro θ es frecuente usar un método pivotal aproximado, que consiste en tomar un estimador de dicho parámetro, $\hat{\theta}$, que una vez estandarizado, converge en distribución a una normal estándar, debido al Teorema Central del Límite. De aquí obtendríamos un intervalo de confianza centrado en el estimador con radio igual al cuantil correspondiente de la normal multiplicado por su error típico.
En nuestro caso de la proporción, se sabe que su estimador es la proporción muestral $\hat{p} = (X_1 + \cdots + X_n)/n$ y sigue aproximadamente una distribución normal de media p y varianza $p(1-p)/n$. Por tanto el intervalo resultante es:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right),$$

El problema de este intervalo es que no se puede calcular, ya que depende del parámetro desconocido p, que se quiere estimar.
Este problema es muy frecuente en los estudios estadísticos de estimación por intervalos de confianza. Veremos en este trabajo que ocurre lo mismo para la media tanto de una distribución de Poisson como de una Exponencial, y también para el coeficiente de correlación.

Una primera solución a este problema será sustituir, en el error típico, el parámetro desconocido por su estimador, obteniendo así un intervalo que en este trabajo lo denominaremos intervalo de confianza *usual*. Aplicando esta primera solución a nuestro ejemplo de la proporción, se obtiene el siguiente intervalo de confianza usual:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

¿Podremos construir un intervalo de confianza mejor? En esto consiste este trabajo en el cual veremos que, aplicando el método delta, se efectúa una transformación al estimador de forma que el error típico ya no dependa del parámetro a estimar. Veremos que al deshacerse dicha transformación se consigue un intervalo de confianza para el parámetro poblacional, que denominaremos intervalo *estabilizado*. Con este tipo de transformación se suele tener una mejor aproximación por la distribución normal y se evita tener que aproximar el error típico. Pues bien, esta transformación es lo que se conoce como *transformación estabilizadora de la varianza* y es el objetivo de estudio del presente trabajo.

Para ello, introduciremos en el primer capítulo los conceptos básicos necesarios para llegar al concepto de intervalo de confianza y al gran Teorema Central del Límite, el cual nos dice a grandes rasgos que sea cual sea la distribución de las X_i variables independientes y con la misma distribución que X, verificando ciertas propiedades, la distribución de la media muestral sigue aproximadamente una distribución normal, cuando el tamaño de la muestra es suficientemente grande.
Llegados a este punto, en el Capítulo 2 enunciaremos el método delta para, a continuación estudiar su aplicación: la transformación estabilizadora de la varianza. Durante estos dos capítulos se pondrá como ejemplo el caso de la proporción mientras que, en la Sección 2.3, veremos con detalle la construcción de los intervalos de confianza por ambos métodos (usual y estabilizado) para los parámetros mencionados anteriormente.
Para finalizar, en el Capítulo 3 se recoge un estudio de simulación que persigue comparar ambos métodos de construcción de intervalos de confianza en cada uno de los ejemplos estudiados. Este estudio se llevará a cabo mediante $M = 10000$ simulaciones de Monte Carlo en las que generaremos M muestras de las distintas distribuciones mencionadas, calcularemos el estimador correspondiente para cada muestra, los intervalos de confianza por ambos métodos y contabilizaremos (de los M) el número de intervalos de confianza que contienen al verdadero valor del parámetro (el cuál conoceremos), esto es la cobertura, que será expresada en porcentaje al igual que los niveles de confianza usuales.
Los códigos empleados para dicha simulación se recogen en el Apéndice A de este trabajo, mientras que a lo largo del Capítulo 3, en sus correspondientes secciones, estarán las tablas y figuras obtenidas por ambos métodos.

Capítulo 1

Aproximación en distribución

1.1. Conceptos básicos

Comenzamos recordando algunos conceptos básicos de la inferencia estadística que serán utilizados en este trabajo, para luego centrarnos en la estimación por intervalos de confianza.

- La *población* es el conjunto de individuos sobre los que se desea conocer ciertas características. Si el interés se centra en una variable aleatoria, es común denotarla por X.

- La *muestra* es un subconjunto observable de la población que sirve para extraer conclusiones sobre toda la población. Es habitual considerar muestras aleatorias simples es decir, variables aleatorias independientes y con la misma distribución de X, que denotaremos por $X_1, \ldots, X_n$. El número de individuos que forma la muestra, n, se denomina *tamaño muestral*.

- El *parámetro*, θ, es una característica de la población, generalmente desconocido y objeto de estudio. Por ejemplo, la media y la varianza son los parámetros en una distribución normal.

- Un *estadístico*, $T(X_1, \ldots, X_n)$, es cualquier función de la muestra que no depende de parámetros desconocidos.

- Un *estimador*, $\hat{\theta}$, es un estadístico que se utiliza para estimar el parámetro de la distribución, θ, verificando ciertas propiedades deseables, como insesgadez (persigue el verdadero valor del parámetro) y consistencia (su variabilidad disminuye al aumentar el tamaño de muestra). Por estar construidos a partir de una muestra aleatoria, los

estimadores son también variables aleatorias. Por lo tanto tienen asociada una cierta distribución, denominada *distribución en el muestreo* y podrán ser utilizados para la construcción de intervalos de confianza, como se verá en la Sección 1.2.2.

1.1.1. Convergencia de sucesiones de variables aleatorias

Sabemos que una *variable aleatoria* es una función de Ω en $\mathbb{R}$, es decir, una función que asocia a cada resultado del espacio muestral un número real. Por otra parte, un *vector aleatorio* es una colección finita de variables aleatorias medidas simultáneamente sobre el mismo individuo o sobre el mismo resultado de un experimento aleatorio. Mientras que una *sucesión de variables aleatorias* es una colección numerable de tales variables aleatorias.

Recordemos las definiciones de los criterios de convergencia de sucesiones de variables aleatorias. Para ello, $\{X_n\}_{n\in\mathbb{N}}$ y X denotarán variables aleatorias definidas sobre el espacio de probabilidad (Ω, A, P).

Definición 1.1. Diremos que la sucesión de variables aleatorias $\{X_n\}_{n\in\mathbb{N}}$ *converge en probabilidad* a X si

$$X_n \xrightarrow{p} X \quad :\Leftrightarrow \quad \forall \varepsilon > 0 \quad \lim_{n\to\infty} P\{|X_n - X| < \varepsilon\} = 1.$$

Definición 1.2. Diremos que la sucesión de variables aleatorias $\{X_n\}_{n\in\mathbb{N}}$ *converge de forma casi segura* a X si

$$X_n \xrightarrow{c.s.} X \quad :\Leftrightarrow \quad P\left\{w \in \Omega / \lim_{n\to\infty} X_n(w) = X(w)\right\} = 1.$$

Definición 1.3. Diremos que la sucesión de variables aleatorias $\{X_n\}_{n\in\mathbb{N}}$ *converge en distribución* a X si

$$X_n \xrightarrow{d} X \quad :\Leftrightarrow \quad \forall x \in C_F \quad \lim_{n\to\infty} F_n(x) = F(x)$$

siendo F_n la función de distribución de $X_n, \forall n$, F la función de distribución de X y C_F el conjunto de puntos de continuidad de F.

Observación 1.4. Se verifican las siguientes relaciones entre los tipos de convergencia:

$$X_n \xrightarrow{c.s.} X \;\Rightarrow\; X_n \xrightarrow{p} X \;\Rightarrow\; X_n \xrightarrow{d} X.$$

Además si X es constante (otorga probabilidad uno a un número real fijo) se verifica

$$X_n \xrightarrow{d} X \;\Rightarrow\; X_n \xrightarrow{p} X.$$

1.2. Estimación

En este punto cabe destacar que en el presente trabajo se realizará una tarea de inferencia estadística desde un punto de vista paramétrico. De esta forma se considerará una población con una cierta función de distribución asociada, $F_\theta(x),\ \theta \in \mathbb{R}^k$, siendo θ el vector de parámetros desconocidos. El objetivo será cuantificar lo mejor posible el valor de θ a partir de una muestra de tamaño n. Para ello existen tres principios metodológicos básicos: la estimación puntual, la estimación por intervalos de confianza y el contraste de hipótesis.

1.2.1. Estimación puntual

Recordemos las propiedades que debe poseer un buen estimador: insesgadez y consistencia.

Definición 1.5. Se denomina *sesgo* de un estimador $\hat{\theta}$ para un parámetro poblacional θ a

$$Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta,$$

y diremos que el estimador es *insesgado* si su sesgo vale cero. Equivalentemente, si $E(\hat{\theta}) = \theta$.

Definición 1.6. Se denomina *error cuadrático medio* de un estimador $\hat{\theta}$ para un parámetro poblacional θ a

$$E\left((\hat{\theta} - \theta)^2\right) = [Sesgo(\hat{\theta})]^2 + Var(\hat{\theta})$$

y diremos que dicho estimador es *consistente* (en media cuadrática) si

$$\lim_{n\to\infty} E\left((\hat{\theta} - \theta)^2\right) = 0.$$

Equivalentemente, si

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta \qquad \text{y} \qquad \lim_{n\to\infty} Var(\hat{\theta}) = 0.$$

Serán preferibles los estimadores con el menor error cuadrático medio posible. En particular, si son insesgados, buscaremos los que tengan la menor varianza posible.

Por este motivo, para estimadores insesgados o con poco sesgo, es habitual presentar cada estimación acompañada de su error típico ($E.T.$) para informar de la calidad de la estimación. Veamos la definición formal.

Definición 1.7. El *error típico* de un estimador $\hat{\theta}$ para un parámetro poblacional θ es su desviación típica:

$$E.T.(\hat{\theta}) = \sqrt{Var(\hat{\theta})}.$$

Sabemos que un buen estimador de un parámetro poblacional (media, proporción, coeficiente de correlación...) va a ser el correspondiente parámetro muestral (media de la muestra, proporción muestral, coeficiente de correlación muestral...). A continuación, a modo ilustrativo, se comprueba la insesgadez y consistencia de la media muestral.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}n\mu = \mu$$

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Por tanto la media muestral es un estimador insesgado (persigue a la media verdadera) y consistente (su precisión aumenta al aumentar el tamaño muestral).

A continuación recodaremos el Teorema de Fisher, que es el instrumento de probabilidad sobre el que se basa la inferencia en poblaciones normales.

Teorema 1.8 (Teorema de Fisher)**.** *Sean $X_1, \ldots, X_n \in N(\mu, \sigma^2)$ independientes. Entonces la media muestral y la varianza o cuasi-varianza muestrales verifican*

(i) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \in N\left(\mu, \frac{\sigma^2}{n}\right)$

(ii) $\frac{nS^2}{\sigma^2} = \frac{(n-1S_c^2)}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \in \chi^2_{n-1}$

(iii) $\bar{X}$ *y* S^2 $(o \quad S_c^2)$ *son independientes.*

Por último ilustremos las propiedades de insesgadez y consistencia con el ejemplo de la proporción muestral.

Ejemplo 1.9. Si queremos obtener información sobre la proporción de individuos con cierta característica en una población, extraeremos una muestra de n individuos donde X_i vale uno si el individuo $i-$ésimo presenta la característica y vale cero en caso contrario. De este modo, $X_1, \ldots, X_n$ son variables independientes e idénticamente distribuidas de $X \in Be(p)$, donde el parámetro p es la proporción poblacional desconocida. El estimador natural de p es la proporción muestral

$$\hat{p} = \frac{X_1 + \cdots + X_n}{n}.$$

Obsérvese que el numerador tiene distribución binomial con n intentos y probabilidad de éxito p, $B(n, p)$ (ya que es la suma de n experimentos independientes de Bernoulli con

probabilidad de éxito p y por tanto tiene media np y varianza $np(1-p)$) y en consecuencia se tiene que

$$E(\hat{p}) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n}np = p,$$

es decir, que $\hat{p}$ es un estimador insesgado de p. Además se tiene que

$$Var(\hat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \overset{n\to\infty}{\longrightarrow} 0,$$

por tanto, al aumentar el tamaño muestral la varianza del estimador tiende a 0. Es decir, $\hat{p}$ es un buen estimador de p ya que es insesgado y consistente. Por último, el error típico de $\hat{p}$ es el siguiente

$$E.T.(\hat{p}) = \sqrt{p(1-p)/n}$$

el cual depende del parámetro desconocido p.

◇

1.2.2. Intervalos de confianza

Como hemos visto, la estimación puntual trata el problema de estimar el valor de una característica poblacional a partir de la información de una muestra utilizando un estimador. Pero en muchos casos esto no es suficiente, ya que no nos aporta información del error que se comete en dicha estimación. Por ello, surge la necesidad de encontrar un método que permita calcular una región que contenga al verdadero valor del parámetro con una cierta probabilidad, es decir un intervalo de confianza. Por tanto el objetivo no es sólo proporcionar el valor estimado obtenido con la muestra, sino una medida de la incertidumbre de dicho valor.

Veremos que para los estimadores simétricos, como es el caso de la media muestral o la proporción, se puede usar como intervalo de confianza

$$\hat{\theta} \pm Cuantil \cdot E.T.(\hat{\theta}).$$

De este modo se obtienen intervalos de confianza centrados en el estimador y cuya amplitud vendrá dada por el cuantil de la correspondiente distribución en el muestreo multiplicado por su error típico.

Cabe observar que cuanto mayor sea la longitud del intervalo, mayor probabilidad de que éste contenga al verdadero valor del parámetro. Pero tampoco buscamos que sea extremadamente grande, ya que, a menor longitud, mayor precisión en la estimación.
En la siguiente definición se recoge de manera formal el concepto de intervalo de confianza.

Definición 1.10. Se llama **intervalo de confianza** para el parámetro θ con nivel de confianza $1-\alpha$, a un intervalo aleatorio (a, b) tal que $P(a < \theta < b) \geq 1-\alpha$. La aleatoriedad del intervalo (a, b) proviene de la aleatoriedad de la muestra, ya que a y b son construidos en base a la muestra.

Si se toman infinitas muestras aleatorias de la población y construimos los correspondientes intervalos de confianza, el $100(1-\alpha)\,\%$ de ellos contendrían el verdadero valor del parámetro, mientras que el $100\alpha\,\%$ no. Como por lo general solo vamos a disponer de una muestra, tenemos que confiar (por ejemplo con un $95\,\%$ de confianza) que la muestra que tenemos pertenece al grupo de las muestras buenas (las que nos dan una estimación del intervalo de confianza que contiene el verdadero valor del parámetro).

Nótese que se habla de nivel de confianza $1-\alpha$, pero la probabilidad es mayor o igual que $1-\alpha$. Esto es porque en muchas situaciones no es posible que dicha probabilidad sea exacta. Para muestras grandes, la probabilidad de que dicho intervalo contenga el valor del parámetro θ es al menos $1-\alpha$ y se le llaman intervalos de confianza de nivel aproximado (o asintótico). En este caso diremos que

$$P(a < \theta < b) \approx 1 - \alpha.$$

¿Por qué calcular intervalos de nivel asintótico?

- Porque no es posible encontrar un pivote cuya distribución no dependa del parámetro.
- Porque no se conoce la distribución exacta del pivote.
- Porque, en general, es más fácil encontrar la distribución asintótica que la exacta del pivote.

A continuación recordaremos qué es un pivote y cómo calcular los extremos del intervalo de confianza por el método pivotal.

1.2.3. Método de la cantidad pivotal

Definición 1.11. Se dice que $T(X_1, \ldots, X_n; \theta)$ es un **pivote** si su distribución es conocida y no depende de θ.

El método de la cantidad pivotal consiste en seleccionar una muestra, escoger el pivote y fijado un nivel de confianza $1-\alpha$ determinar las constantes a y b tales que

$$P(a \leq T(X_1, \ldots, X_n; \theta) \leq b) = 1 - \alpha.$$

Si es posible despejar θ de la expresión anterior, obtendremos dos valores que determinan el intervalo, es decir

$$P\left(T^{-1}(X_1,\ldots,X_n;a) \leq \theta \leq T^{-1}(X_1,\ldots,X_n;b)\right) = 1-\alpha.$$

Veámoslo con el ejemplo para la media de una distribución normal, ya que posteriormente veremos el por qué de su importancia.

Ejemplo 1.12. Sea $X_1,\ldots,X_n$ una muestra de variables aleatorias independientes y con la misma distribución de $X \in N(\mu,\sigma^2)$ con varianza σ^2 conocida. Hallemos un intervalo de confianza de nivel $1-\alpha$ para la media por el método pivotal. Por el Teorema de Fisher 1.8 sabemos que $\overline{X} \in N\left(\mu, \frac{\sigma^2}{n}\right)$. Tipificando resulta que

$$T(X_1,\ldots,X_n;\mu) = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \backsim N(0,1),$$

que puede ser utilizado como pivote ya que su distribución no depende de μ.
Por tanto, llamando $z_{\alpha/2}$ al cuantil que deja probabilidad $\alpha/2$ a la derecha en una distribución normal estándar, se puede obtener un intervalo de confianza para la media, μ, sin más que despejarla de la siguiente expresión

$$P\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}\right) = 1-\alpha.$$

Haciendo cálculos para despejar μ, resulta que

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

y por tanto,

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{1.1}$$

es un intervalo de confianza para μ con nivel de confianza $1-\alpha$. Por comodidad también se denota como $\left(\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$.

$\diamond$

En este ejemplo se consideró el caso de varianza conocida, pero esto es poco realista. Cuando la varianza es desconocida ha de ser estimada usando la cuasivarianza (S_c) y en este caso la distribución del pivote sería una T-Student con $n-1$ grados de libertad dando lugar al siguiente intervalo de confianza:

$$\left(\overline{X} \pm t_{n-1,\alpha/2}\frac{S_c}{\sqrt{n}}\right),$$

donde $t_{n-1,\alpha/2}$ es el cuantil $\alpha/2$ de la distribución T_{n-1}.

La distribución T-Student se puede aproximar por una $N(0,1)$ cuando el número de grados de libertad es grande. En este caso se podría reemplazar el cuantil de la T-Student por el cuantil de la normal estándar y obtener un intervalo aproximado.

¿Cómo elegir el pivote? Se buscarán funciones sencillas de la muestra y del parámetro, cuya distribución no dependa del parámetro θ. En los casos más sencillos, el pivote surge de forma natural.

Observación 1.13. En algunas situaciones la búsqueda es más complicada. En estos casos, si tenemos muestras grandes y se verifican ciertas condiciones, podemos recurrir como pivote al estimador de máxima verosimilitud $\hat{\theta}_{MV}$, que es asintóticamente normal, obteniéndose como pivote

$$\frac{\hat{\theta}_{MV} - \theta}{\sqrt{(I(\theta))^{-1/2}}} \xrightarrow{d} N(0,1),$$

donde $I(\theta) = -E\left(\frac{d^2 \log f(x,\theta)}{d\theta^2}\right)$. No nos detendremos más en esta cuestión. Para más información se puede consultar la página 68 de [2].

Pero, ¿qué sucede cuando la población no sigue una distribución normal? En la sección siguiente estudiaremos la teoría asintótica que se emplea para aproximar la distribución del estimador (o del pivote) aun cuando los datos no proceden de la normal.

1.3. Teorema central del límite.

Hemos visto que para construir un intervalo de confianza para la media de una población, el estimador natural es la media muestral $\overline{X}$, pero puede suceder que se desconozca su distribución y por tanto no se pueda construir dicho intervalo. Para superar esta dificultad y construir intervalos de confianza asintóticos, se utiliza el teorema central del límite el cual nos asegura que sea cual sea la distribución de las variables X_i, cumpliendo ciertas propiedades, la distribución de la media muestral es aproximadamente una normal, siempre que el tamaño de la muestra sea grande.

Veremos la formulación general del teorema central del límite y enunciaremos la versión que nos interesa, la de Lévy-Lindeberg.

Teorema 1.14 (Teorema central del límite)**.** *Diremos que una sucesión de variables aleatorias $\{X_n\}_{n\in\mathbb{N}}$ satisface el Teorema central del límite con límite Y si existen sucesiones $\{A_n\}_{n\in\mathbb{N}}(A_n \in \mathbb{R})$, y otra de números positivos $\{B_n\}_{n\in\mathbb{N}}$ tal que $\lim_{n\to\infty} B_n = \infty$ verificando que*

$$\frac{\sum_{k=1}^{n} X_k - A_n}{B_n} \xrightarrow{d} Y.$$

Esto da lugar al siguiente resultado, de gran utilidad para la Estadística, tanto a nivel teórico como práctico, ya que está en términos de variables aleatorias independientes e idénticamente distribuidas.

Teorema 1.15 (Teorema central del límite de Lévy-Lindeberg)**.** *Sea* $\{X_n\}_{n\in\mathbb{N}}$ *una sucesión de variables aleatorias tal que* X_i *son independientes e idénticamente distribuidas a* X *con media y varianzas finitas* $\mu = E(X_i), \sigma^2 = Var(X_i), \forall i$ *entonces*

$$\frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}(X_k - \mu) \xrightarrow{d} N(0,1). \tag{1.2}$$

Dividiendo la expresión (1.2) por n, resulta que

$$\frac{\frac{\sum_{k=1}^{n} X_k}{n} - \mu}{\frac{\sigma\sqrt{n}}{n}} = \sqrt{n}\ \frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{d} N(0,1), \tag{1.3}$$

donde se ve una analogía con la tipificación de una variable aleatoria.

Esta última expresión (1.3) nos permite obtener un pivote con el cual podemos construir un intervalo de confianza asintótico para la media poblacional (suponiendo σ^2 conocida) de la misma forma que en el Ejemplo 1.12, resultando así el intervalo

$$\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{1.4}$$

Al intervalo para la media (1.4) lo denominaremos **usual**, ya que en los estudios de simulación que serán realizados en el Capítulo 3 se comparará con los intervalos que estudiaremos en el Capítulo 2. También veremos con el ejemplo para la proporción que cuando σ depende del parámetro será necesario añadir una corrección.

Otras expresiones equivalentes de gran utilidad obtenidas al reescribir la expresión (1.3) son:

1. La distribución de la media muestral se puede aproximar por una normal:

$$\overline{X}_n \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right) \tag{1.5}$$

2. En el Capítulo 2 resultará de gran utilidad el siguiente enunciado:

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2). \tag{1.6}$$

En definitiva, el estudio de la distribución de un estimador para un n finito es a menudo complejo y, en ocasiones involucra cálculos prácticamente imposibles de llevar a cabo. Por

este motivo, en muchas ocasiones, solo es posible realizar estudios del comportamiento asintótico (cuando n tiende a infinito) de los estimadores. En la práctica, la gran mayoría de los estimadores usuales, centrados y normalizados, tienen distribución asintótica normal resultado que se obtiene del teorema central del límite.

Observación 1.16. Si la muestra es pequeña, $\overline{X}_n$ tendrá una distribución en el muestreo que no será normal y por tanto, no tendremos ninguna garantía de que $\left(\overline{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ tenga el nivel de confianza deseado.

Además, notemos que si X es normal, $\overline{X}_n$ es normal para cualquier tamaño muestral y por tanto, el intervalo de confianza será exacto, no asintótico.

Como curiosidad, el adjetivo *central* del teorema central del límite es debido a Polyá (1920), y significa fundamental, o de importancia central. Esto manifiesta la grandiosidad de este teorema en la inferencia estadística y desvela las razones por las cuales, en muchos campos de aplicación, se encuentran en todo momento distribuciones normales, o asintóticamente normales.

A continuación enunciaremos la versión multivariante del Teorema 1.15.

Teorema 1.17 (Teorema de Lévy-Lindeberg multivariante)**.** *Si tenemos una sucesión de vectores aleatorios independientes y con la misma distribución de un vector X, con vector de medias y matriz de covarianzas finitos $\mu = E(X)$ y $\Sigma = Cov(X, X)$, entonces*

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(X_k - \mu) \xrightarrow{d} N(0, \Sigma).$$

1.3.1. Aplicación a la proporción poblacional

Por último en este capítulo aplicaremos el Teorema 1.15 al caso de una variable con distribución de Bernoulli con el objetivo de buscar un intervalo de confianza para la proporción poblacional.

Ejemplo 1.18. Sean $X_1, \ldots, X_n$ una muestra aleatoria independiente y con la misma distribución de $X \in Be(p)$, y por tanto $E(X_i) = p \quad \text{y} \quad Var(X_i) = p(1-p)$ $i = 1, \ldots, n$. En el Ejemplo 1.9 vimos que un buen estimador para la proporción poblacional es la proporción muestral $\hat{p} = \frac{X_1 + \cdots + X_n}{n}$.
Por el teorema central del límite tenemos que

$$\sqrt{n}\frac{\hat{p} - p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0, 1),$$

por tanto tenemos un pivote (ya que su distribución no depende de p) y podemos construir un intervalo de confianza para la media de forma análoga al Ejemplo 1.12:

$$1-\alpha \approx P\left(\hat{p} - z_{\alpha/2}\sqrt{p(1-p)/n} < p < \hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n}\right),$$

por tanto el intervalo de confianza asintótico resultante para la proporción es

$$\left(\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right).$$

Pero observemos que el $E.T.(\hat{p})$ depende del parámetro desconocido p y por tanto no se puede calcular. Una primera opción es estimar el error típico, sustituyendo p por su estimador, la proporción muestral $\hat{p}$, dando como resultado el siguiente intervalo de confianza

$$\left(\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right), \tag{1.7}$$

que tiene la forma *usual* que buscábamos en (1.4) y que usaremos para los estudios de simulación.

◇

Nótese que en este intervalo usual, los extremos pueden caer fuera del intervalo $(0, 1)$ y veremos que este problema estará presente en el intervalo estabilizado del Capítulo 2.

Capítulo 2

Método de estabilización de varianza

En este capítulo presentamos el método de estabilización de la varianza y lo aplicamos a ciertos ejemplos notables. La herramienta de probabilidad que da pie a la estabilización de la varianza es el método delta, que estudiamos en la Sección 2.1.

2.1. Método delta

Empezamos recordando dos teoremas importantes relacionados con sucesiones de variables aleatorias y definiendo formalmente qué es una sucesión de estimadores asintóticamente normal y consistente.

Teorema 2.1 (Teorema de Slutsky). *Sean* $\{X_n\}_{n\in\mathbb{N}}$ *e* $\{Y_n\}_{n\in\mathbb{N}}$ *sucesiones de variables aleatorias, sea* X *una variable aleatoria y sea* c *una constante tales que* $X_n \xrightarrow{d} X$ *e* $Y_n \xrightarrow{p} c$, *entonces*

1. $X_n \pm Y_n \xrightarrow{d} X \pm c$
2. $\begin{cases} X_n Y_n \xrightarrow{d} Xc & si\ c \neq 0 \\ X_n Y_n \xrightarrow{p} 0 & si\ c = 0 \end{cases}$
3. $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}\ si\ c \neq 0$

Teorema 2.2 (Teorema de la Aplicación Continua). *Sean* $\{X_n\}_{n\in\mathbb{N}}$ *una sucesión de variables aleatorias tal que* $X_n \xrightarrow{p} X$ *y* g *una función continua en el rango de valores de* X, *entonces*

$$g(X_n) \xrightarrow{p} g(X).$$

Recordemos que este teorema también se cumple para la convergencia en distribución.

Definición 2.3. Una sucesión de estimadores $T_n = \{T_n\}_{n\in\mathbb{N}}$, se dice *asintóticamente normal y consistente para* θ si existe una sucesión $\sigma_n(\theta) > 0$ tal que $\lim_{n\to\infty} \sigma_n(\theta) = 0 \ \forall\theta \in \Theta$ y

$$\frac{T_n - \theta}{\sigma_n(\theta)} \xrightarrow{d} N(0,1) \ \forall\theta \in \Theta$$

De esta definición podríamos obtener un intervalo de confianza para θ pero como hemos visto no podría calcularse debido a que $\sigma_n(\theta)$ depende del parámetro desconocido. Por el Teorema 2.2 sabemos que si $T_n \xrightarrow{p} \theta$ y g es una función contínua en θ, entonces $g(T_n) \xrightarrow{p} g(\theta)$. Ahora bien, si $\sqrt{n}(T_n - \theta)$ es asintóticamente normal $N(0,\sigma^2)$ ¿podremos saber si $\sqrt{n}(g(T_n) - g(\theta))$ también es asintóticamente normal? El método delta se encarga de responder a esta pregunta, siendo la respuesta afirmativa si g es diferenciable.

Teorema 2.4 (Método delta)**.** *Si* $\{T_n\}_{n\in\mathbb{N}}$ *es una sucesión de estimadores de* θ *tal que*

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0,\sigma^2)$$

y $g : \mathbb{R} \longrightarrow \mathbb{R}$ *es una función diferenciable en* θ *con* $g'(\theta) \neq 0$*, entonces*

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N(0, g'(\theta)^2\sigma^2). \tag{2.1}$$

Como en el caso del teorema central del límite, veamos la extensión del método delta al caso multidimensional.

Teorema 2.5 (Método delta multidimensional)**.** *Si* $\{T_n\}_{n\in\mathbb{N}}$ *es una sucesión de vectores aleatorios de dimensión* k *tales que*

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N_k(0,\Sigma)$$

y $g : \mathbb{R}^k \to \mathbb{R}^m$ *es una función diferenciable en* θ*, siendo* $Dg(\theta)$ *la matriz jacobiana de* g *en* θ*, entonces*

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N_m\left(0, Dg(\theta)\Sigma Dg(\theta)^t\right)$$

siempre que $Dg(\theta) \neq 0$.

2.2. Transformación estabilizadora de la varianza

Hemos visto que, si $\{T_n\}_{n\in\mathbb{N}}$ es una sucesión de variables aleatorias verificando que $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0,\sigma^2(\theta))$, un intervalo de confianza asintótico para θ con nivel de confianza aproximado $1-\alpha$ viene dado por

$$\left(T_n - z_{\alpha/2}\frac{\sigma(\theta)}{\sqrt{n}}, T_n + z_{\alpha/2}\frac{\sigma(\theta)}{\sqrt{n}}\right). \tag{2.2}$$

Como ya se ha mencionado, el problema de estos intervalos es que la varianza asintótica depende del parámetro desconocido θ y una primera solución es reemplazar la desviación típica por un estimador. Si dicho estimador se elige consistente, el intervalo de confianza resultante seguirá teniendo un nivel de confianza asintótico $1-\alpha$.

Otra solución es usar una transformación que estabilice la varianza, que a menudo nos conduce a una mejor aproximación. El método delta nos decía que si g es diferenciable, entonces

$$\sqrt{n}(g(T_n)-g(\theta)) \xrightarrow{d} N\left(0, g'(\theta)^2\sigma^2(\theta)\right).$$

Si elegimos g tal que $g'(\theta)^2\sigma^2(\theta)=k^2$, siendo k^2 constante, entonces k^2 sería la varianza asintótica, que no dependería del parámetro desconocido θ, ya que tendríamos que

$$\sqrt{n}(g(T_n)-g(\theta)) \xrightarrow{d} N(0,k^2)$$

y de este modo encontrar un intervalo de confianza asintótico para $g(\theta)$ sería fácil.

Definición 2.6. Supongamos que $k=1$, despejando $g(\theta)$ de la expresión $g'(\theta)^2\sigma^2(\theta)=1$ resulta que $g'(\theta)=\frac{1}{\sigma(\theta)}$ y por tanto se obtiene:

$$g(\theta)=\int \frac{1}{\sigma(\theta)}d\theta.$$

A esta solución se le denomina **transformación estabilizadora de la varianza.**

Si g está bien definida, un intervalo de confianza asintótico para $g(\theta)$ será

$$\left(g(T_n) \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\right). \tag{2.3}$$

Deshaciendo el cambio, tenemos un nuevo tipo de intervalo de confianza para θ, ya que

$$P\left(g^{-1}\left(g(T_n)-z_{\alpha/2}\frac{1}{\sqrt{n}}\right) \leq \theta \leq g^{-1}\left(g(T_n)+z_{\alpha/2}\frac{1}{\sqrt{n}}\right)\right) \approx 1-\alpha. \tag{2.4}$$

2.2.1. Aplicación a la proporción poblacional

Apliquemos el método de la transformación estabilizadora de la varianza a nuestro ejemplo de la proporción:

Ejemplo 2.7. En el Ejemplo 1.18 vimos que

$$\sqrt{n}\frac{\hat{p}-p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1) \Leftrightarrow \sqrt{n}(\hat{p}-p) \xrightarrow{d} N(0,p(1-p)).$$

Por el método delta resulta que

$$\sqrt{n}(g(\hat{p})-g(p)) \xrightarrow{d} N(0,g'(p)^2p(1-p)). \tag{2.5}$$

Ahora, eligiendo g tal que $g'(p)\sqrt{p(1-p)} = 1$, tenemos que $g'(p) = \frac{1}{\sqrt{p(1-p)}}$ e integrando resulta que

$$g(p) = \int \frac{1}{\sqrt{p(1-p)}} dp.$$

Haciendo el cambio de variable $u^2 = p$, $2udu = dp$, resulta que la integral anterior es equivalente a la siguiente integral

$$2\int \frac{1}{\sqrt{1-u^2}} du = 2 \text{ arcsen}(u)$$

Deshaciendo el cambio, tenemos entonces que

$$g(p) = 2 \text{ arcsen}(\sqrt{p})$$

es una transformación estabilizadora de la varianza y sustituyendo en la expresión (2.5), resulta

$$\sqrt{n}(2 \text{ arcsen}(\hat{p}) - 2 \text{ arcsen}(\sqrt{p})) \xrightarrow{d} N\left(0, \frac{1}{p(1-p)} p(1-p)\right) = N(0,1).$$

Es decir, tomando como pivote

$$T(X_1, \ldots, X_n; 2 \text{ arcsen}(\hat{p})) = \frac{2 \text{ arcsen}(\sqrt{\hat{p}}) - 2 \text{ arcsen}(\sqrt{p})}{\frac{1}{\sqrt{n}}} \Rightarrow$$

$$P\left(-z_{\alpha/2} \leq \frac{2 \text{ arcsen}(\sqrt{\hat{p}}) - 2 \text{ arcsen}(\sqrt{p})}{\frac{1}{\sqrt{n}}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

y despejando 2 arcsen$(\sqrt{p})$ tenemos el intervalo de confianza asintótico para 2 arcsen$(\sqrt{p})$

$$\left(2 \text{ arcsen}(\sqrt{\hat{p}}) \pm \frac{1}{\sqrt{n}} z_{\alpha/2}\right).$$

Deshaciendo el cambio como indicamos en la expresión (2.4), tenemos que:

$$\begin{aligned} 1 - \alpha &\approx P\left(-\text{arcsen}(\sqrt{\hat{p}}) - \frac{z_{\alpha/2}}{2\sqrt{n}} \leq -\text{arcsen}(\sqrt{p}) \leq -\text{ arcsen}(\sqrt{\hat{p}}) + \frac{z_{\alpha/2}}{2\sqrt{n}}\right) = \\ &= P\left(\text{sen}\left(\text{arcsen}(\sqrt{\hat{p}}) - \frac{z_{\alpha/2}}{2\sqrt{n}}\right) \leq \sqrt{p} \leq \text{sen}\left(\text{arcsen}(\sqrt{\hat{p}}) + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right) = \\ &= P\left(\text{sen}^2\left(\text{arcsen}(\sqrt{\hat{p}}) - \frac{z_{\alpha/2}}{2\sqrt{n}}\right) \leq p \leq \text{sen}^2\left(\text{arcsen}(\sqrt{\hat{p}}) + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right). \end{aligned}$$

Obsérvese en el extremo inferior del intervalo, que arcsen$(\sqrt{\hat{p}})$ toma valores entre 0 y $\pi/2$ pero el arcsen$(\sqrt{\hat{p}}) - z_{\alpha/2}/(2\sqrt{n})$ puede ser negativo y arcsen$(\sqrt{\hat{p}}) + z_{\alpha/2}/(2\sqrt{n})$ puede ser mayor que $\pi/2$. Como $\sqrt{p}$ es siempre positivo, reemplazamos el extremo inferior por 0 si resulta ser negativo y por 1 si resulta ser mayor que 1. Por tanto el intervalo de confianza asintótico para p con la varianza estabilizada es:

$$\left(\text{sen}^2\left(\text{máx}\left(\text{arcsen}(\sqrt{\hat{p}}) - \frac{z_{\alpha/2}}{2\sqrt{n}}\right), 0\right), \text{sen}^2\left(\text{mín}\left(\text{arcsen}(\sqrt{\hat{p}}) + \frac{z_{\alpha/2}}{2\sqrt{n}}\right), 1\right)\right). \quad (2.6)$$

$\diamond$

2.3. Ejemplos notables

A continuación repasaremos brevemente ciertas distribuciones notables, incluida la distribución de Bernoulli con la que hemos estudiado los distintos conceptos de este trabajo, para seguidamente construir los dos tipos de intervalos de confianza: el usual y el que estabiliza la varianza, que compararemos por medio de un estudio de simulación en el siguiente capítulo.

2.3.1. Distribución de Bernoulli

Los experimentos de Bernoulli son aquellos que sólo presentan dos posibles resultados: éxito y fracaso, por tanto la variable X toma dos valores: 1 con probabilidad de éxito p, y 0 con probabilidad $1-p$.
Obsérvese que al ser p una probabilidad, esta se encontrará entre 0 y 1.

Si X es una variable aleatoria que mide el número de éxitos, y se realiza un único experimento con dos posibles resultados (éxito o fracaso), se dice que la variable aleatoria X se distribuye como una Bernoulli de parámetro p y se denota por $X \in Be(p)$. Recordemos sus principales características y los dos tipos de intervalos que hemos calculado:

1. La fórmula de la función masa de probabilidad de la distribución de Bernouilli es:
   $f(x) = p^x(1-p)^{1-x}, \qquad x = \{0, 1\}$

2. $E(X) = p$

3. $Var(X) = p(1-p)$

En el Ejemplo 1.18 obtuvimos la expresión (1.7) para el intervalo de confianza asintótico usual para la proporción p. El intervalo resultante era

$$\left(\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Mientras que, con la transformación que estabiliza la varianza, en el Ejemplo 2.7 obteníamos en (2.6) el siguiente intervalo:

$$\left(\text{sen}^2\left(\text{máx}\left(\text{arcsen}(\sqrt{\hat{p}}) - \frac{z_{\alpha/2}}{2\sqrt{n}}\right), 0\right), \text{sen}^2\left(\text{mín}\left(\text{arcsen}(\sqrt{\hat{p}}) + \frac{z_{\alpha/2}}{2\sqrt{n}}\right), 1\right)\right).$$

2.3.2. Distribución de Poisson

Un proceso de Poisson, que se denota por $Pois(\lambda)$, consiste en observar el número de veces que se presenta un suceso (número de éxitos) en un determinado intervalo (generalmente de tiempo). En estos procesos se asume que hay estabilidad, en el sentido de que el número de sucesos por unidad de tiempo, λ, permanece constante.

1. La fórmula de la función masa de probabilidad de la distribución de Poisson es: $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots$

2. $E(X) = Var(X) = \lambda$.

3. Además puede comprobarse que $\hat{\lambda}_n = \overline{X}_n$ coincide con el estimador de máxima verosimilitud y también con el obtenido por el método de los momentos.

Ahora, sea $X_1, \ldots, X_n$ una muestra aleatoria de observaciones independientes y con la misma distribución de $X \in Pois(\lambda)$. Queremos construir un intervalo de confianza asintótico para la media por el método usual. Para ello aplicando la expresión (1.3) obtenida del teorema central del límite se tiene que

$$\sqrt{n}\frac{\overline{X}_n - \lambda}{\sqrt{\lambda}} \xrightarrow{d} N(0,1) \tag{2.7}$$

y por los razonamientos usuales, tenemos el siguiente intervalo de confianza:

$$\left(\overline{X}_n \pm z_{\alpha/2}\sqrt{\frac{\lambda}{n}}\right),$$

el cual depende del parámetro desconocido λ, que habría que estimar. Por coincidir la media poblacional con el parámetro de la Poisson, lo natural es sustituir λ en el error típico por el estimador dado por la media muestral $\overline{X}_n$, dando como resultado el intervalo de confianza asintótico usual buscado,

$$\left(\overline{X}_n \pm z_{\alpha/2}\sqrt{\frac{\overline{X}_n}{n}}\right). \tag{2.8}$$

Busquemos ahora el intervalo de confianza asintótico para la media, pero esta vez, usando una transformación estabilizadora de la varianza. Para ello aplicaremos el método delta. Como se cumple, por la expresión (2.7), que $\sqrt{n}(\overline{X}_n - \lambda) \xrightarrow{d} N(0, \lambda)$, se verifica también que

$$\sqrt{n}(g(\overline{X}_n) - g(\lambda)) \xrightarrow{d} N\left(0, g'(\lambda)^2\sigma^2(\lambda)\right). \tag{2.9}$$

Elegimos g diferenciable tal que $g'(\lambda)\sigma(\lambda) = 1$, y por tanto, $g'(\lambda) = \frac{1}{\sigma(\lambda)} = \frac{1}{\sqrt{\lambda}}$.
Integrando resulta que

$$g(\lambda) = \int \frac{1}{\sqrt{\lambda}} d\lambda = \int \lambda^{-1/2} d\lambda = 2\sqrt{\lambda}$$

es una transformación estabilizadora de la varianza, y así sustituyendo en la expresión (2.9) tenemos

$$\sqrt{n}\left(2\sqrt{\overline{X}_n} - 2\sqrt{\lambda}\right) \xrightarrow{d} N\left(0, \lambda\frac{1}{\lambda}\right) = N(0,1).$$

Es decir, tomando como pivote

$$T(X_1, \ldots, X_n; 2\sqrt{\lambda}) = \frac{2\sqrt{\overline{X}_n} - 2\sqrt{\lambda}}{\frac{1}{\sqrt{n}}} \Rightarrow P\left(-z_{\alpha/2} \leq \frac{2\sqrt{\overline{X}_n} - 2\sqrt{\lambda}}{\frac{1}{\sqrt{n}}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

y despejando $2\sqrt{\lambda}$ tenemos un intervalo de confianza asintótico para $2\sqrt{\lambda}$ con nivel de confianza aproximado $1 - \alpha$ que es

$$\left(2\sqrt{\overline{X}_n} \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\right).$$

Deshaciendo el cambio como indicamos en la expresión (2.4), tenemos que

$$2\sqrt{\overline{X}_n} - z_{\alpha/2}\frac{1}{\sqrt{n}} \leq 2\sqrt{\lambda} \leq 2\sqrt{\overline{X}_n} + z_{\alpha/2}\frac{1}{\sqrt{n}} \Rightarrow$$

$$\Rightarrow \left(\sqrt{\overline{X}_n} - z_{\alpha/2}\frac{1}{2\sqrt{n}}\right)^2 \leq \lambda \leq \left(\sqrt{\overline{X}_n} + z_{\alpha/2}\frac{1}{2\sqrt{n}}\right)^2.$$

Por tanto, haciendo cuentas, el intervalo de confianza asintótico para la media con la varianza estabilizada es

$$\left(\overline{X}_n + \frac{1}{4n}z_{\alpha/2}^2 \pm \sqrt{\frac{\overline{X}_n}{n}}z_{\alpha/2}\right). \tag{2.10}$$

2.3.3. Distribución exponencial

Hemos visto que un proceso de Poisson se utilizaba para medir el número de sucesos de un determinado tipo que tenían lugar en un determinado intervalo. Consideramos ahora la variable X que estudia el tiempo entre dos sucesos consecutivos. Esta seguirá una distribución exponencial de parámetro λ y se denota por $Exp(\lambda)$. Podemos decir entonces que una distribución de Poisson mide el número de sucesos en un intervalo de tiempo, y una exponencial el tiempo que tarda en producirse un suceso partiendo de un instante determinado.

1. La función de densidad de la distribución exponencial es

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{en otro caso.} \end{cases}$$

2. $E(X) = \frac{1}{\lambda}$

3. $Var(X) = \frac{1}{\lambda^2}$

Ahora, sea $X_1, \ldots, X_n$ una muestra aleatoria de observaciones independientes y con la misma distribución de $X \in Exp(\lambda)$ con la que queremos construir un intervalo de confianza asintótico para la media por el método usual. Para ello, aplicando el teorema central del límite (1.3) se tiene que

$$\sqrt{n}\frac{\overline{X}_n - \theta}{\theta} \xrightarrow{d} N(0,1) \tag{2.11}$$

y por los razonamientos usuales, tenemos el siguiente intervalo de confianza:

$$\left(\overline{X}_n \pm z_{\alpha/2}\frac{\theta}{\sqrt{n}}\right), \tag{2.12}$$

el cual depende del parámetro desconocido θ, que habría que estimar. Lo natural es sustituir, en el error típico, θ por su estimador: la media muestral $\overline{X}_n$, dando como resultado el intervalo de confianza asintótico usual buscado

$$\left(\overline{X}_n \pm z_{\alpha/2}\frac{\overline{X}_n}{\sqrt{n}}\right). \tag{2.13}$$

El intervalo de confianza asintótico para la media usando una transformación estabilizadora de la varianza se puede obtener teniendo en cuenta que se cumple por la expresión (2.11) $\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{d} N(0, \theta^2)$ Por tanto, se verifica que

$$\sqrt{n}(g(\overline{X}_n) - g(\theta)) \xrightarrow{d} N\left(0, g'(\theta)^2\sigma^2(\theta)\right). \tag{2.14}$$

Buscamos g diferenciable tal que $g'(\theta)\sigma(\theta) = 1$, y por tanto $g'(\theta) = \frac{1}{\sigma(\theta)} = \frac{1}{\theta}$. Integrando resulta que

$$g(\theta) = \int \frac{1}{\theta} d\theta = \ln\theta,$$

es una transformación estabilizadora de la varianza. Sustituyendo en la expresión (2.14) tenemos

$$\sqrt{n}\left(\ln\overline{X}_n - \ln\theta\right) \xrightarrow{d} N\left(0, \theta^2\frac{1}{\theta^2}\right) = N(0,1).$$

Es decir, tomando como pivote

$$T(X_1, \ldots, X_n; \ln\theta) = \frac{\ln\overline{X}_n - \ln\theta}{\frac{1}{\sqrt{n}}} \Rightarrow P\left(-z_{\alpha/2} \leq \frac{\ln\overline{X}_n - \ln\theta}{\frac{1}{\sqrt{n}}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

y despejando $\ln\theta$ tenemos el intervalo de confianza asintótico para $\ln\theta$ con nivel de confianza aproximado $1-\alpha$

$$\left(\ln \overline{X}_n \pm \frac{z_{\alpha/2}}{\sqrt{n}}\right).$$

Deshaciendo el cambio como indicamos en la expresión (2.4), resulta

$$\overline{X}_n e^{-z_{\alpha/2}/\sqrt{n}} \leq \theta \leq \overline{X}_n e^{z_{\alpha/2}/\sqrt{n}},$$

así, el intervalo de confianza asintótico para la media con la varianza estabilizada es

$$\left(\overline{X}_n e^{-z_{\alpha/2}/\sqrt{n}}, \overline{X}_n e^{z_{\alpha/2}/\sqrt{n}}\right). \tag{2.15}$$

2.3.4. Coeficiente de correlación

Ahora queremos construir un intervalo de confianza para el coeficiente de correlación. Para ello consideramos $(X_1, Y_1), \ldots, (X_n, Y_n)$ una muestra aleatoria de observaciones independientes del vector (X, Y) que sigue una distribución normal bivariante, es decir

$$(X,Y)' \sim N\left(\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix}\right),$$

y con coeficiente de correlación poblacional

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Se toma como estimador de ρ el coeficiente de correlación muestral definido por

$$r_n = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}$$

Para distribuciones generales para las cuales exista el cuarto momento, se verifica que $\sqrt{n}(r-\rho)$ tiene una distribución aproximadamente normal de media cero y varianza que depende del tercer y cuarto momento. Asumiendo la normalidad, la varianza asintótica se puede expresar en términos de la correlación de X e Y. Tomamos como cierto el siguiente resultado, cuya demostración omitiremos (véase [3]):

$$\sqrt{n}(r_n - \rho) \xrightarrow{d} N(0, (1-\rho^2)^2).$$

Por tanto, se verifica que

$$\sqrt{n}\frac{r_n - \rho}{1-\rho^2} \xrightarrow{d} N(0,1),$$

y por los razonamientos usuales, tenemos el siguiente intervalo de confianza

$$\left(r_n \pm z_{\alpha/2}\frac{(1-\rho^2)}{\sqrt{n}}\right),$$

que depende del parámetro desconocido ρ. De nuevo, sustituyendo en el error típico ρ por su estimador natural, el coeficiente de correlación muestral, r_n, obtenemos el intervalo de confianza asintótico usual buscado

$$\left(r_n \pm z_{\alpha/2}\frac{(1-r_n^2)}{\sqrt{n}}\right). \tag{2.16}$$

Obtengamos ahora el intervalo de confianza asintótico para el coeficiente de correlación, pero esta vez, usando una transformación estabilizadora de la varianza. Como se cumple $\sqrt{n}(r_n-\rho) \sim N(0,(1-\rho^2)^2)$, usando el método delta tendríamos que

$$\sqrt{n}(g(r_n)-g(\rho)) \xrightarrow{d} N\left(0, g'(\rho)^2\sigma^2(\rho)\right). \tag{2.17}$$

Elegimos g diferenciable tal que $g'(\rho)\sigma(\rho)=1$, y por tanto $g'(\rho)=\frac{1}{\sigma(\rho)}=\frac{1}{1-\rho^2}$. Integrando resulta que

$$g(\rho)=\int \frac{1}{1-\rho^2}d\rho = \int\left(\frac{1}{2(1-\rho)}+\frac{1}{2(1+\rho)}\right)d\rho = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) = \operatorname{arctanh}\,\rho$$

es una transformación estabilizadora de la varianza y sustituyendo los datos en la expresión (2.17) tenemos

$$\sqrt{n}(\operatorname{arctanh}\, r_n - \operatorname{arctanh}\,\rho) \xrightarrow{d} N\left(0, \frac{1}{(1-\rho^2)^2}(1-\rho^2)^2\right) = N(0,1).$$

Es decir, tomando como pivote

$$T(X_1,\ldots,X_n) = \frac{\operatorname{arctanh}\, r_n - \operatorname{arctanh}\,\rho}{\frac{1}{\sqrt{n}}} \Rightarrow P\left(-z_{\alpha/2} \leq \frac{\operatorname{arctanh}\, r_n - \operatorname{arctanh}\,\rho}{\frac{1}{\sqrt{n}}} \leq z_{\alpha/2}\right) \approx 1-\alpha$$

y despejando arctanh ρ tenemos un intervalo de confianza asintótico para arctanh ρ con nivel de confianza aproximado $1-\alpha$ que es

$$\left(\operatorname{arctanh}\, r_n \pm \frac{z_{\alpha/2}}{\sqrt{n}}\right).$$

Deshaciendo el cambio como indicamos en la expresión (2.4) tenemos que

$$\operatorname{arctanh}\, r_n - \frac{z_{\alpha/2}}{\sqrt{n}} \leq \operatorname{arctanh}\,\rho \leq \operatorname{arctanh}\, r_n + \frac{z_{\alpha/2}}{\sqrt{n}}.$$

Por último, despejando ρ, el intervalo de confianza asintótico para el coeficiente de correlación poblacional con la varianza estabilizada es

$$\left(\tanh\left(\operatorname{arctanh}\, r_n - \frac{z_{\alpha/2}}{\sqrt{n}}\right), \tanh\left(\operatorname{arctanh}\, r_n + \frac{z_{\alpha/2}}{\sqrt{n}}\right)\right). \tag{2.18}$$

Capítulo 3

Estudios de simulación

En este capítulo haremos un estudio de simulación para comparar la cobertura de los dos tipos de intervalos asintóticos calculados en el capítulo anterior: los usuales y los que estabilizan la varianza. Este estudio se hará a través de una simulación basada en $M = 10000$ simulaciones de Monte Carlo, para ello nos situaremos en un escenario controlado donde conoceremos los verdaderos valores de los parámetros poblacionales y así podremos evaluar la cobertura obtenida en los distintos intervalos de confianza asintóticos.

Además de presentar los resultados de la cobertura de los intervalos de confianza construidos por los dos métodos, se añadirán algunas representaciones gráficas de las distribuciones de las estimaciones obtenidas antes y después de la transformación.

Este estudio se hará mediante el programa estadístico `R` [4]. Los códigos utilizados para la obtención de los resultados se recogen en el Apéndice A, mientras que las tablas y representaciones gráficas se recogen en las secciones correspondientes a cada distribución. Como se puede observar en el código que aparece en el Apéndice A se ha fijado la semilla al principio de cada simulación, lo cual garantiza que el experimento es reproducible por cualquier otro investigador.

3.1. Distribución de Bernoulli

Siguiendo el orden de este trabajo, empezaremos por el caso de la media de una población que sigue una distribución de Bernoulli. Por tanto, sabemos que la media coincide con la probabilidad de éxito p.

Por ser la distribución de Bernoulli discreta, la media muestral tomará un número finito de valores, por tanto lo más apropiado es representar la distribución de la media muestral mediante un gráfico de barras.

Empezamos analizando los resultados obtenidos en la Tabla 3.1 correspondientes a muestras de tamaño $n = 15$. Lo deseable es que la cobertura coincida con el nivel de confianza indicado en cada columna (90 %, 95 % y 99 % respectivamente). En general, la mejor aproximación a la verdadera confianza del intervalo se tiene para valores de p en torno a 0.5 para todos los niveles de confianza. Atribuimos este hecho a la simetría de la distribución binomial para $p = 0.5$, simetría que se pierde gradualmente conforme p se acerca a 0 o a 1, empeorando así la aproximación por la distribución normal.

Puede verse que para $p = 0.1$ la cobertura del intervalo usual es mejor que la cobertura del intervalo estabilizado, pero en ambos casos se obtienen valores muy por debajo del nivel de confianza. Por tanto, para valores de p próximos a 0 (y también próximos a 1) ninguno de los dos métodos parece ser adecuado.

Es importante destacar que la cobertura para p, fijado un nivel de confianza, coincide con la cobertura de su complementario $1 - p$ y esto se verifica para ambos intervalos y para los distintos tamaños muestrales. Esto es debido a la forma del error típico, en el cual aparece $p(1 - p)$ y coincide con su complementario. Entonces en este caso, cuando hablemos de la cobertura para algún valor p, se entenderá que ocurre lo mismo para su complementario. Por este motivo en las Tablas 3.2 y 3.3, correspondientes a tamaños muestrales $n = 25$ y $n = 50$, respectivamente, se muestran las coberturas para valores de $p \leq 0.5$.

Se puede ver que en ocasiones coinciden las coberturas del intervalo usual y el estabilizado. Esto es debido al carácter discreto de la distribución binomial y el reducido tamaño de muestra.

En general, tanto el intervalo usual como el intervalo con estabilización de la varianza presentan coberturas inferiores al nivel de confianza.

Las coberturas del intervalo con estabilización de la varianza son, en general, mejores que las correspondientes al intervalo usual, especialmente para el nivel del 99 %. Por ejemplo, para $p = 0.25$ y $p = 0.4$ al 99 % de confianza siempre se ve una mejor cobertura para los intervalos estabilizados que para los usuales.

p	90 %	95 %	99 %
0.1	77.71	78.89	79.06
0.25	87.43	91.10	92.33
0.4	88.09	94.22	96.55
0.5	88.61	88.61	96.69
0.6	88.09	94.22	96.55
0.75	87.43	91.10	92.33
0.9	77.71	78.89	79.06

p	90 %	95 %	99 %
0.1	73.47	77.71	78.89
0.25	87.43	91.10	98.44
0.4	88.09	94.22	98.68
0.5	88.61	96.69	99.19
0.6	88.09	94.22	98.68
0.75	87.43	91.10	98.44
0.9	73.47	77.71	78.89

Tabla 3.1: Cobertura de los intervalos usuales (izquierda) y estabilizados (derecha) para la media de una distribución de Be(p) para muestras de tamaño $n = 15$.

p	90 %	95 %	99 %
0.1	91.74	91.74	92.59
0.25	88.46	90.37	96.80
0.4	89.71	94.32	98.76
0.5	89.81	95.93	98.76

p	90 %	95 %	99 %
0.1	89.37	91.74	92.42
0.25	88.46	94.43	99.17
0.4	89.71	94.32	98.76
0.5	89.81	95.93	98.76

Tabla 3.2: Cobertura de los intervalos usuales (izquierda) y estabilizados (derecha) para la media de una distribución de Be(p) para muestras de tamaño $n = 25$.

p	90 %	95 %	99 %
0.1	86.06	87.60	96.35
0.25	87.78	93.95	97.94
0.4	89.18	94.03	98.01
0.5	88.50	93.61	98.63

p	90 %	95 %	99 %
0.1	82.73	93.87	99.15
0.25	89.99	92.50	99.10
0.4	89.18	94.03	99.13
0.5	88.50	93.61	98.63

Tabla 3.3: Cobertura de los intervalos usuales (izquierda) y estabilizados (derecha) para la media de una distribución de Be(p) para muestras de tamaño $n = 50$.

A continuación analizamos las representaciones gráficas (Figuras 3.1, 3.2 y 3.3) de la proporción muestral y su transformación estabilizadora. Se trata de diagramas de barras de frecuencias relativas para algunos valores de p y n. En el eje horizontal se representan los valores de la media muestral (figura de la izquierda) y del arcoseno de la raíz cuadrada de la media muestral (figura de la derecha). En el eje vertical se representan las frecuencias relativas en las 10000 muestras simuladas.

Podemos ver que para $p = 0.25$ y muestras de tamaño pequeño, la distribución es muy asimétrica en el caso de la proporción muestral, mientras que para la transformación mediante el arcoseno de su raíz cuadrada esta asimetría se corrige. Este efecto sigue siendo visible a medida que aumenta el tamaño de la muestra.
Lo mismo ocurre, aunque es menos notable, para valores de la proporción próximos a 0.5, esto puede verse en la Figura 3.4, donde el valor de la verdadera proporción es $p = 0.4$.

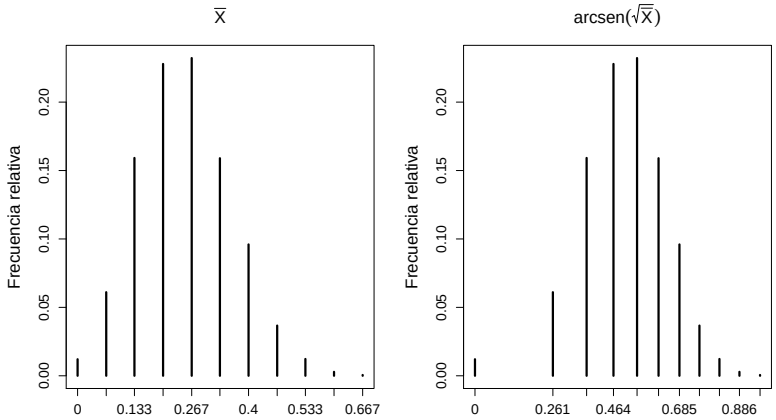

Figura 3.1: Diagrama de barras de la distribución de la media muestral(izquierda) y del arcoseno de la raíz cuadrada de la media muestral (derecha) para $p = 0.25$ y $n = 15$.

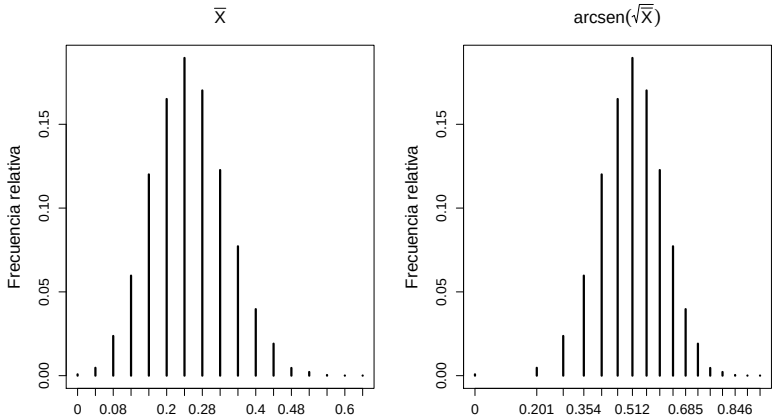

Figura 3.2: Diagrama de barras de la distribución de la media muestral(izquierda) y del arcoseno de la raíz cuadrada de la media muestral (derecha) para $p = 0.25$ y $n = 25$.

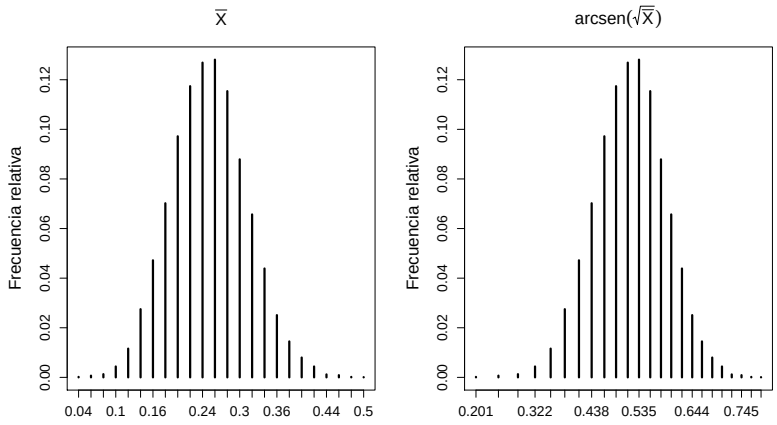

Figura 3.3: Diagrama de barras de la distribución de la media muestral(izquierda) y del arcoseno de la raíz cuadrada de la media muestral (derecha) para $p = 0.25$ y $n = 50$.

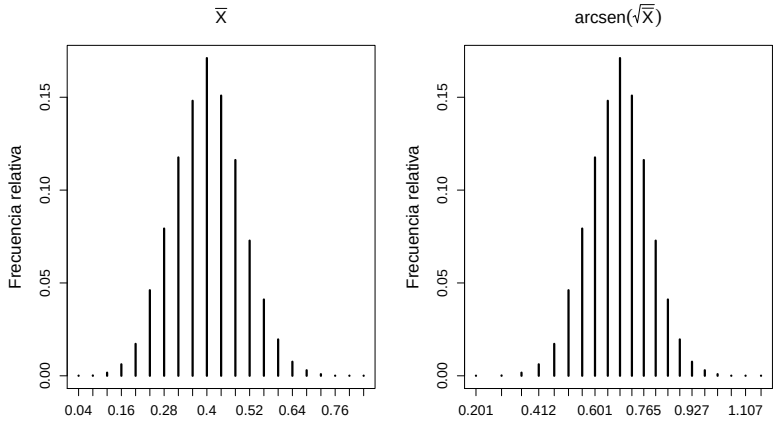

Figura 3.4: Diagrama de barras de la distribución de la media muestral(izquierda) y del arcoseno de la raíz cuadrada de la media muestral (derecha) para $p = 0.4$ y $n = 25$.

3.2. Distribución de Poisson

Veamos el caso de la media de una población que sigue una distribución de Poisson, donde la media coincide con el parámetro λ.

Puede verse en las Tablas 3.4, 3.5 y 3.6 correspondientes a los tamaños de muestra $n = 15, n = 25$ y $n = 50$ respectivamente que las coberturas obtenidas, en ambos métodos

de construcción de intervalos, con valor de $\lambda = 0.25$ resultan ser las peores. De hecho, generalmente (con estos datos), las peores coberturas estimadas ocurren para valores del verdadero parámetro λ menores que 1 siendo, en algunos casos, incluso peores las obtenidas en los intervalos estabilizados (derecha) que las obtenidas en los intervalos usuales (izquierda). Puede verse que a medida que aumentamos el tamaño muestral, n, esto mejora. Fijémonos por ejemplo, en la Tabla 3.6 correspondiente a $n = 50$ se tiene que para $\lambda = 0.25$ con un 90 % de nivel de confianza, es mejor la cobertura obtenida para los intervalos usuales que para los estabilizados, pero ya no es así a medida que aumentamos el nivel de confianza.

En general, para los distintos $\lambda > 1$ se obtienen buenas coberturas siendo mejores las obtenidas en los intervalos estabilizados (derecha) y mejorando ambos a medida que se aumenta el tamaño muestral n.

En este caso haremos dos representaciones gráficas ya que por ser la distribución de Poisson discreta, cuando el valor de λ es pequeño, la media muestral tomará un número finito de valores, por tanto representaremos la distribución de la media muestral a partir de diagramas de barras.

Por otro lado, cuando el valor de λ es grande, $\lambda \geq 10$, sabemos como consecuencia del teorema central del límite 1.15 que la distribución de Poisson se puede aproximar por una normal de media y varianza igual a λ. De esta manera, podemos considerar la distribución de la media muestral continua y analizar gráficamente su distribución, mediante el histograma de las estimaciones a lo largo de las M simulaciones y su estimación de la densidad. Además, para estudiar lo buena que es la aproximación a la distribución normal en cada caso, se añadirá la curva de la densidad normal teórica.

En las representaciones mediante diagramas de barras (Figuras 3.5 y 3.6) se aprecia la asimetría de la distribución de la media muestral antes y después de la transformación, propiedad que mejora cuando el verdadero parámetro λ se hace más grande. Como hemos explicado, para valores de λ grande, la distribución de Poisson se aproxima a una normal y puede apreciarse en la Figura 3.7 correspondiente al tamaño de muestra $n = 15$ que, tanto antes como después de la transformación raíz cuadrada de la media muestral, hay una buena aproximación a la distribución normal teórica. En cuanto a la simetría de la distribución de la media muestral, no se aprecia ninguna mejora después de la transformación.

λ	90 %	95 %	99 %
0.25	88.31	89.24	89.64
0.75	91.34	92.70	96.78
1	90.43	92.33	98.30
2	89.02	93.30	98.74
5	89.22	95.45	98.95
10	89.69	95.34	99.06
25	89.59	94.74	99.07

λ	90 %	95 %	99 %
0.25	86.20	96.36	97.60
0.75	89.64	94.95	98.52
1	90.43	94.58	99.03
2	90.41	94.48	99.00
5	89.76	95.22	98.87
10	90.00	95.04	99.01
25	89.70	94.46	99.11

Tabla 3.4: Cobertura de los intervalos usuales (izquierda) y estabilizados (derecha) para la media de una distribución de Poisson(λ) para muestras de tamaño $n = 15$.

λ	90 %	95 %	99 %
0.25	84.78	94.58	94.99
0.75	90.93	95.27	97.92
1	87.90	94.92	97.81
2	90.64	95.14	98.88
5	90.57	95.27	98.87
10	89.87	95.15	99.05
25	90.00	94.95	98.93

λ	90 %	95 %	99 %
0.25	92.59	93.93	98.62
0.75	89.18	94.73	98.81
1	89.32	94.12	99.19
2	89.57	94.63	99.02
5	90.57	94.80	99.18
10	90.25	95.08	99.02
25	89.92	94.89	98.97

Tabla 3.5: Cobertura de los intervalos usuales (izquierda) y estabilizados (derecha) para la media de una distribución de Poisson(λ) para muestras de tamaño $n = 25$.

λ	90 %	95 %	99 %
0.25	89.71	92.04	96.48
0.75	90.53	93.87	98.74
1	90.10	94.81	98.74
2	89.84	94.47	98.76
5	89.60	94.95	98.94
10	89.99	94.80	99.16
25	90.25	95.10	98.96

λ	90 %	95 %	99 %
0.25	87.73	94.63	98.33
0.75	89.38	94.87	98.81
1	89.12	94.17	98.83
2	89.84	94.69	98.91
5	89.81	94.85	98.95
10	90.14	95.02	99.11
25	90.48	95.05	98.95

Tabla 3.6: Cobertura de los intervalos usuales (izquierda) y estabilizados (derecha) para la media de una distribución de Poisson(λ) para muestras de tamaño $n = 50$.

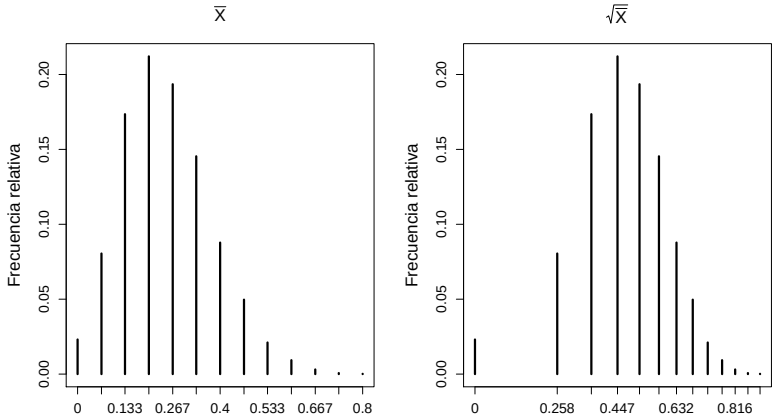

Figura 3.5: Diagrama de barras de la distribución de la media muestral(izquierda) y de la raíz cuadrada de la media muestral (derecha) para $\lambda = 0.25$ y $n = 15$.

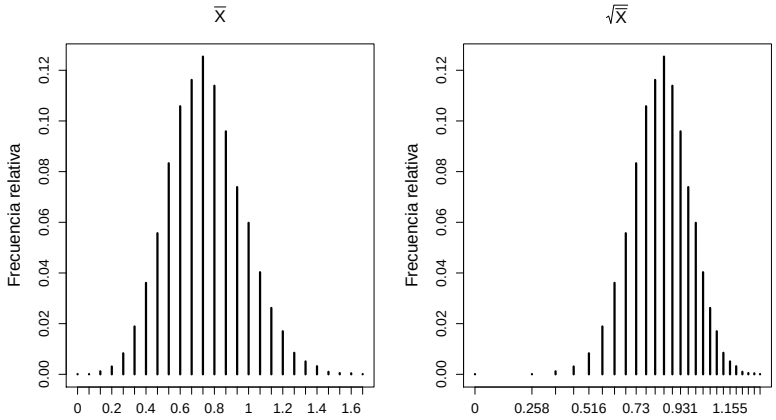

Figura 3.6: Diagrama de barras de la distribución de la media muestral (izquierda) y de la raíz cuadrada de la media muestral (derecha) para $\lambda = 0.75$ y $n = 15$.

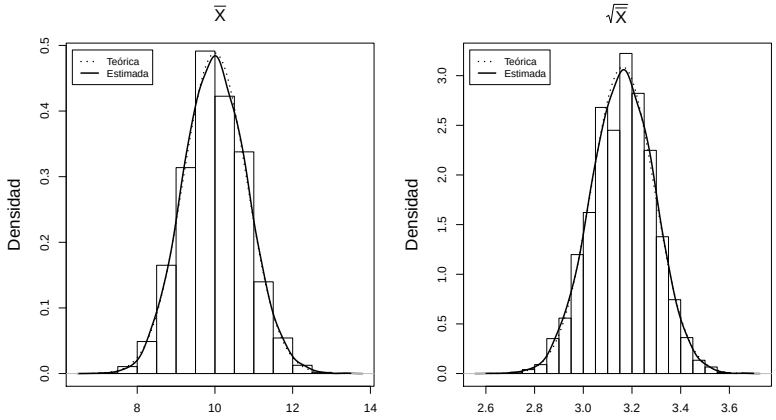

Figura 3.7: Histograma de la distribución de la media muestral (izquierda) y de la raíz cuadrada de la media muestral (derecha) para $\lambda = 10$ y $n = 15$.

3.3. Distribución Exponencial

Veamos el caso de la media de una población que sigue una distribución exponencial. Por tanto, sabemos que la media coincide con la inversa del parámetro de la distribución Exponencial. Por ser la distribución Exponencial continua representaremos la distribución de la media muestral, mediante histogramas y estimación de la densidad. Además se añadirá la curva de la densidad normal teórica.

Cabe destacar que en este caso se tiene la misma cobertura para cualquiera que sea el valor del parámetro λ. De esta manera, en la Tabla 3.7 se recogen los porcentajes de cobertura para un único valor $\lambda = 1$, distinguiendo las coberturas de los intervalos usuales de los estabilizados, y para distintos tamaños muestrales.

n	90 %	95 %	99 %
15	88.29	91.84	95.84
25	88.38	92.99	97.08
50	89.18	93.88	97.74

n	90 %	95 %	99 %
15	89.55	94.43	98.48
25	89.43	94.76	98.75
50	89.60	94.63	98.99

Tabla 3.7: Cobertura de los intervalos usuales (izquierda) y estabilizados (derecha) para la media de una distribución Exp(1) para muestras de distintos tamaños muestrales, n.

Observando la Tabla 3.7 de la izquierda, es decir las coberturas de los intervalos de confianza usuales, puede apreciarse que a medida que aumentamos el tamaño muestral n, aumentan los porcentajes de cobertura acercándose cada vez más al nivel de confianza, como cabía esperar, ya que la convergencia en distribución a la normal será mejor a medida que aumente el tamaño muestral, según vimos en el teorema central del límite 1.15. Ocurre lo mismo para los intervalos de confianza con la transformación estabilizadora de la varianza, donde el porcentaje de cobertura prácticamente coincide con el nivel de confianza. Aún así podemos apreciar que todas las coberturas de los intervalos estabilizados (derecha) son más próximas a la cobertura teórica que las obtenidas en los intervalos usuales (izquierda) y por tanto, son mejores los intervalos de confianza estabilizados, cosa que perseguimos a lo largo de este trabajo y en este caso se cumple a la perfección.

En la Figura 3.8, correspondiente al tamaño muestral $n = 50$, puede verse el beneficio de aplicar la transformación logaritmo a la media muestral, ya que se tiene una simetrización de la distribución.

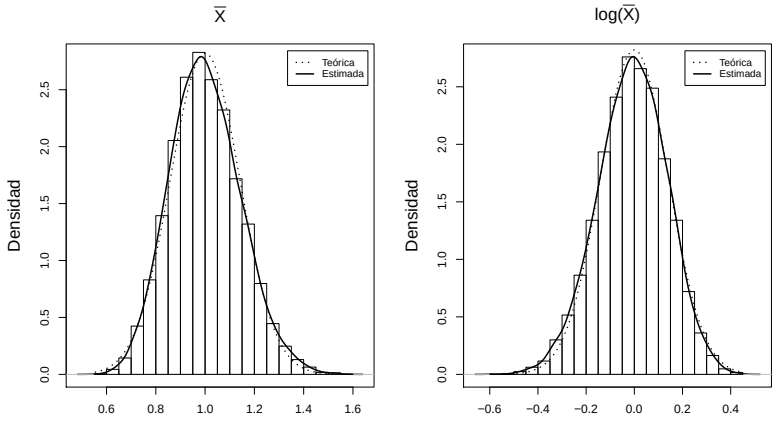

Figura 3.8: Histogramas de la distribución de la media muestral(izquierda) y del logaritmo de la media muestral (derecha) para $\lambda = 1$ y $n = 50$.

3.4. Coeficiente de correlación

Para finalizar, consideraremos el caso de estimar el coeficiente de correlación ρ de una población normal bivariante. Recordemos brevemente que para estudiar la relación entre dos variables se definen la covarianza y el coeficiente de correlación. A diferencia de la covarianza, el coeficiente de correlación es una medida estandarizada y adimensional que toma valores entre -1 y 1. Se verifica que si las variables son independientes, la covarianza es igual a cero, y por tanto también el coeficiente de correlación es nulo. En este caso se dice que las variables son incorrelacionadas. Generalmente, el recíproco no es cierto aunque sí se verifica si las variables siguen una distribución normal.

En este caso simularemos dos variables normales estándar con distintas correlaciones fijas y distintos tamaños muestrales ($n = 15, n = 25$ y $n = 50$). Se calculará el coeficiente de correlación muestral para cada una de las $M = 10000$ muestras y se obtiene el número de intervalos construidos por cada método que contienen al verdadero valor del parámetro, obteniendo así la cobertura. Aunque ρ toma valores entre -1 y 1, no consideraremos los valores negativos ya que, como en el caso de la proporción, se obtiene la misma cobertura que para su correspondiente valor positivo.

Si nos fijamos en la Tabla 3.8 correspondiente al tamaño de muestra $n = 15$, puede comprobarse que para todos los valores del verdadero coeficiente de correlación, se obtie-

ne una mejor cobertura en los intervalos obtenidos por el método de la transformación estabilizadora de la varianza (derecha) que en los intervalos usuales (izquierda).

Analicemos qué ocurre si aumentamos el tamaño muestral: en las Tablas 3.9 y 3.10 correspondientes al tamaño de muestras $n = 25$ y $n = 50$ respectivamente, podemos ver que la cobertura mejora en ambos métodos y se sigue teniendo una mejor aproximación de las coberturas en los intervalos estabilizados.

ρ	90 %	95 %	99 %
0	82.07	88.08	94.34
0.1	82.56	88.06	94.26
0.25	82.77	88.01	94.10
0.5	83.04	87.95	93.61
0.75	83.07	87.74	92.77
0.9	83.35	87.49	92.03

ρ	90 %	95 %	99 %
0	86.19	92.25	97.67
0.1	86.15	92.34	97.53
0.25	86.14	92.35	97.54
0.5	86.28	92.31	97.63
0.75	86.54	92.23	97.65
0.9	86.64	92.25	97.69

Tabla 3.8: Cobertura de los intervalos usuales (izquierda) y estabilizados (derecha) para distintos valores del verdadero coeficiente de correlación ρ para muestras de tamaño $n = 15$.

ρ	90 %	95 %	99 %
0	85.43	90.69	96.43
0.1	85.24	91.01	96.42
0.25	85.37	91.09	96.30
0.5	85.37	90.92	96.13
0.75	85.83	90.43	95.39
0.9	85.76	90.09	94.84

ρ	90 %	95 %	99 %
0	87.83	93.19	98.36
0.1	88.06	93.41	98.41
0.25	87.99	93.43	98.43
0.5	87.98	93.54	98.44
0.75	88.01	93.63	98.50
0.9	87.98	93.68	98.52

Tabla 3.9: Cobertura de los intervalos usuales (izquierda) y estabilizados (derecha) para distintos valores del verdadero coeficiente de correlación ρ para muestras de tamaño $n = 25$.

ρ	90 %	95 %	99 %
0	87.29	92.78	97.85
0.1	87.69	92.75	97.71
0.25	87.71	92.67	97.62
0.5	87.63	92.68	97.47
0.75	87.74	92.53	96.90
0.9	87.66	92.24	96.63

ρ	90 %	95 %	99 %
0	88.66	94.15	98.53
0.1	88.88	94.01	98.63
0.25	88.91	94.05	98.61
0.5	88.92	94.13	98.59
0.75	89.00	94.13	98.58
0.9	88.98	94.14	98.58

Tabla 3.10: Cobertura de los intervalos usuales (izquierda) y estabilizados (derecha) para distintos valores del verdadero coeficiente de correlación ρ para muestras de tamaño $n = 50$.

Para finalizar este trabajo, observando la Figura 3.9, correspondiente a un tamaño muestral $n = 15$, puede apreciarse que cuando hay mucha correlación entre las variables, la distribución del coeficiente de correlación muestral (izquierda) es muy asimétrica, cosa que mejora después de aplicar la transformación arcotangente hiperbólico a dicho coeficiente de correlación. Además se obtiene una mejor aproximación a la distribución teórica normal después de aplicar dicha transformación, la cual va mejorando a medida que aumentamos el tamaño muestral como cabía esperar (véase la Figura 3.10).

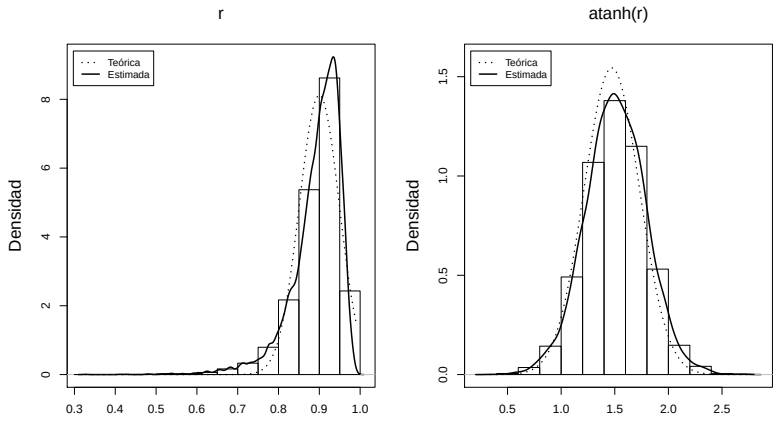

Figura 3.9: Histogramas de la distribución del coeficiente de correlación muestral (izquierda) y del arcotangente hiperbólico del coeficiente de correlación muestral (derecha) para $\rho = 0.9$ y $n = 15$.

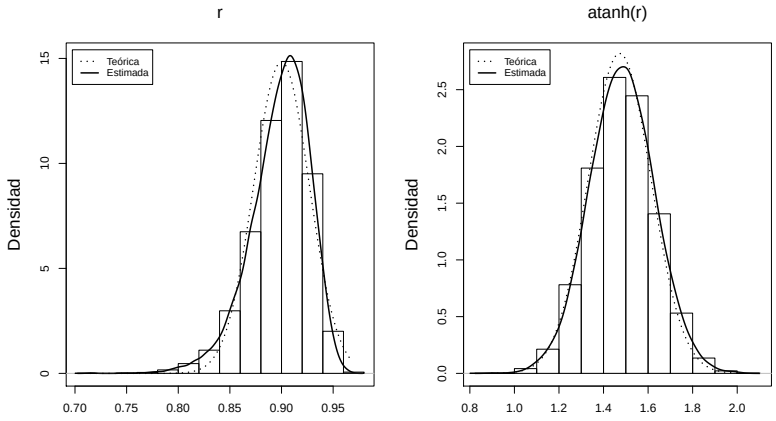

Figura 3.10: Histogramas de la distribución del coeficiente de correlación muestral (izquierda) y del arcotangente hiperbólico del coeficiente de correlación muestral (derecha) para $\rho = 0.9$ y $n = 50$.

Como conclusión final podemos afirmar que, en general, la estabilización de la varianza aporta una mejora clara en distribuciones continuas pero para distribuciones discretas, su efecto no es tan nítido.

Apéndice A : Código de R

Bernoulli

```
n=15                                        # tamaño de muestra: 15, 25, 50
ns=10000                                    # número de simulaciones
vmedia=c(0.1,0.25,0.4,0.5,0.6,0.75,0.9)     # distintos parámetros
lparam=length(vmedia)                       # longitud vector de parámetros
nivel=c(0.90,0.95,0.99)                     # distintos niveles de confianza
alfa=1-nivel
z=qnorm(1-alfa/2)                           # cuantil de la normal estándar

# Reserva memoria cobertura
c1=matrix(0,ncol=length(nivel),nrow=lparam)
c2=matrix(0,ncol=length(nivel),nrow=lparam)
rownames(c1)=vmedia;colnames(c1)=nivel
rownames(c2)=vmedia;colnames(c2)=nivel

# Reserva memoria almacenamiento de las estimaciones
vmed=matrix(0,nrow=ns,ncol=lparam)

for (iparam in 1:lparam){
  set.seed(123456)
  media=vmedia[iparam]
  p=media

  for (is in 1:ns){
    x=rbinom(n,size=1,prob=p)
    med=mean(x)                             # media muestral
    vmed[is,iparam]=med                     # almacenar
```

```
    #-- Método usual
    et=sqrt(med*(1-med)/n)
    inf1=med-z*et
    sup1=med+z*et
    c1[iparam,] = c1[iparam,] + ((inf1<media)&(media<sup1))

    #-- Método estabilizado
    aux1=asin(sqrt(med))
    aux2=z/(2*sqrt(n))
    inf2=pmax(sin(aux1-aux2)^2,0)
    sup2=sin(aux1+aux2)^2
    c2[iparam,] = c2[iparam,] + ((inf2<media)&(media<sup2))
    }
}

# Cobertura en porcentaje
#-- Método usual
c1/ns*100
#-- Método estabilizado
c2/ns*100
```

Poisson

```
n=15                                    # tamaño de muestra: 15, 25, 50
ns=10000                                # número de simulaciones
vmedia=c(0.25,0.75,1,2,5,10,25)         # distintos parámetros
lparam=length(vmedia)                   # longitud vector de parámetros
nivel=c(0.90,0.95,0.99)                 # distintos niveles de confianza
alfa=1-nivel
z=qnorm(1-alfa/2)                       # cuantil de la normal estándar

# Reserva memoria cobertura
c1=matrix(0,ncol=length(nivel),nrow=lparam)
c2=matrix(0,ncol=length(nivel),nrow=lparam)
rownames(c1)=vmedia;colnames(c1)=nivel
```

```
rownames(c2)=vmedia;colnames(c2)=nivel

# Reserva memoria almacenamiento de las estimaciones
vmed=matrix(0,nrow=ns,ncol=lparam)

for (iparam in 1:lparam){
  set.seed(123456)
  media=vmedia[iparam]
  lambda=media

  for (is in 1:ns){
    x=rpois(n,lambda)
    med=mean(x)                          # media muestral
    vmed[is,iparam]=med                  # almacenar

    #-- Método usual
    et=sqrt(med/n)
    inf1=med-z*et
    sup1=med+z*et
    c1[iparam,] = c1[iparam,] + ((inf1<media)&(media<sup1))

    #-- Método estabilizado
    inf2=med+z^2*(1/(4*n))-z*sqrt(med/n)
    sup2=med+z^2*(1/(4*n))+z*sqrt(med/n)
    c2[iparam,] = c2[iparam,] + ((inf2<media)&(media<sup2))
    }
}

# Cobertura en porcentaje
#-- Método usual
c1/ns*100
#-- Método estabilizado
c2/ns*100
```

Exponencial

```
n=15                            # tamaño de muestra: 15, 25, 50
ns=10000                        # número de simulaciones
vmedia=c(0.5,1,2,4,8)           # distintos parámetros
lparam=length(vmedia)           # longitud vector de parámetros
nivel=c(0.90,0.95,0.99)         # distintos niveles de confianza
alfa=1-nivel
z=qnorm(1-alfa/2)               # cuantil de la normal estándar

# Reserva memoria cobertura
c1=matrix(0,ncol=length(nivel),nrow=lparam)
c2=matrix(0,ncol=length(nivel),nrow=lparam)
rownames(c1)=vmedia;colnames(c1)=nivel
rownames(c2)=vmedia;colnames(c2)=nivel

# Reserva memoria almacenamiento de las estimaciones
vmed=matrix(0,nrow=ns,ncol=lparam)

for (iparam in 1:lparam){
  set.seed(123456)
  media=vmedia[iparam]
  lambda=1/media

  for (is in 1:ns){
    x=rexp(n,rate=lambda)
    med=mean(x)                 # media muestral
    vmed[is,iparam]=med         # almacenar

    #-- Método usual
    et=med/sqrt(n)
    inf1=med-z*et
    sup1=med+z*et
    c1[iparam,] = c1[iparam,] + ((inf1<media)&(media<sup1))

    #-- Método estabilizado
    inf2=med*exp(-z/sqrt(n))
```

```
    sup2=med*exp(z/sqrt(n))
    c2[iparam,] = c2[iparam,] + ((inf2<media)&(media<sup2))
    }
}

# Cobertura en porcentaje
#-- Método usual
c1/ns*100
#-- Método estabilizado
c2/ns*100
```

Coeficiente de correlación

```
library(MASS)                          # librería para simular normales multidimensionales
n=15                                   # tamaño de muestra: 15, 25, 50
ns=10000                               # número de simulaciones
rhos=c(0,0.1,0.25,0.5,0.75,0.9)        # distintos parámetros
lpar=length(rhos)                      # longitud vector de parámetros
nivel=c(0.90,0.95,0.99)                # distintos niveles de confianza
alfa=1-nivel
z=qnorm(1-alfa/2)                      # cuantil de la normal estándar

# Reserva memoria cobertura
c1=matrix(0,ncol=length(nivel),nrow=lpar)
c2=matrix(0,ncol=length(nivel),nrow=lpar)
rownames(c1)=rhos; colnames(c1)=nivel
rownames(c2)=rhos; colnames(c2)=nivel

# Reserva memoria almacenamiento de las estimaciones
corr_mat=matrix(0,nrow=ns,ncol=lpar)

for (ipar in 1:lpar){
  set.seed(654321)
  rho=rhos[ipar]

  for (is in 1:ns){
```

```
    data=mvrnorm(n, mu=c(0, 0), Sigma=matrix(c(1, rho, rho, 1), nrow=2))
    x=data[,1]                          # normal estándar
    y=data[,2]                          # normal estándar
    r_n=cor(x,y)                        # coeficiente correlación muestral
    corr_mat[is,ipar]=r_n               # almacenar

    #-- Método usual
    et=(1-r_n^2)/sqrt(n)
    inf1=r_n-z*et
    sup1=r_n+z*et
    c1[ipar,]=c1[ipar,]+((inf1<rho)&(rho<sup1))

    #-- Método estabilizado
    at=atanh(r_n)
    inf2=tanh(at-z/sqrt(n))
    sup2=tanh(at+z/sqrt(n))
    c2[ipar,]=c2[ipar,]+((inf2<rho)&(rho<sup2))
    }
}

# Cobertura en porcentaje
#-- Método usual
c1/ns*100
#-- Método estabilizado
c2/ns*100
```

Bibliografía

[1] *https* : *//www.ine.es/explica/docs/historia_estadistica.pdf*

[2] Espejo Miranda, I. Fernández Palacín, F. López Sánchez, M. A. Muñoz Márquez, M. Rodriguez Chía, A. M. Sánchez Navas, A. y Valero Franco, C. (2007) *Inferencia Estadística*, Servicio de Publicaciones de la Universidad de Cádiz.

[3] Van der Vaart, A.W. (1998). *Asymptotic statistics.* Cambridge University Press.

[4] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL *http* : *//www.R-project.org/*.